Homework

January 26, 2021

0.0.1 Homework 1, Erika Ergart

Question 16.1

In order to assess the impact of September 11, BTS took the following approach: using data before September 11, they forecasted future data (under the assumption of no terrorist attack). Then, they compared the forecasted series with the actual data to assess the impact of the event. Our first step, therefore, is to split each of the time series into two parts: pre- and post September 11. We now concentrate only on the earlier time series.

a). Is the goal of this study descriptive or predictive?

The goal of the study is predictive because it was assumed that there is no terrorist attack on 9/11, which fails to describe the reality (or to be a descriptive model)

b. Plot each of the three pre-event time series (Air, Rail, Car).

[1]:
```
library(forecast)
```

```
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

[3]:
```
colnames(df)
```

1. 'Month' 2. 'Air.RPM..000s.' 3. 'Rail.PM' 4. 'VMT..billions.'

[4]:
```
names(df)[2] <- "Air"
names(df)[3] <- "Rail"
names(df)[4] <- "Car"
```

[5]:
```
head(df,2)
```

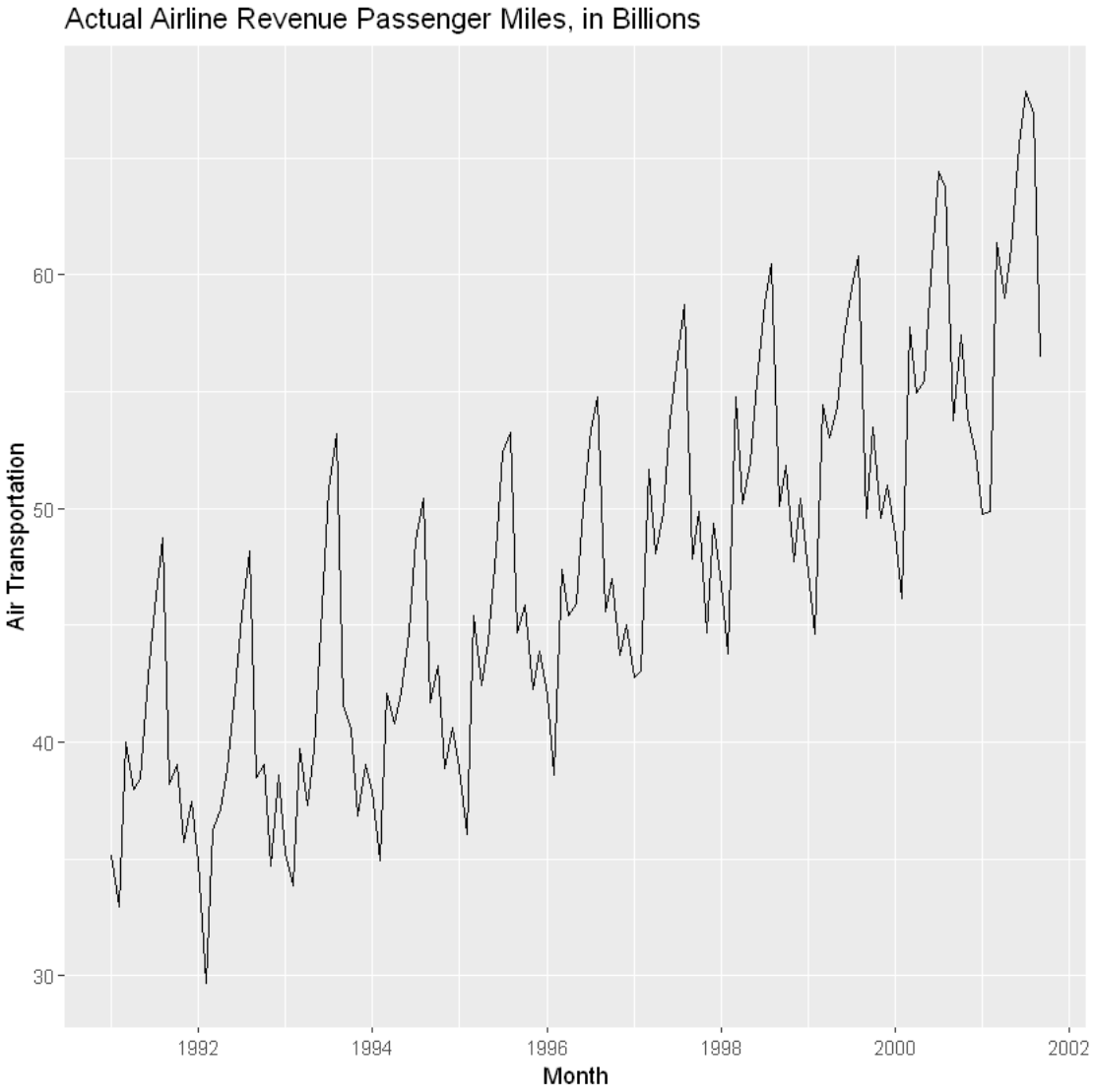
A data.frame: 2 × 4

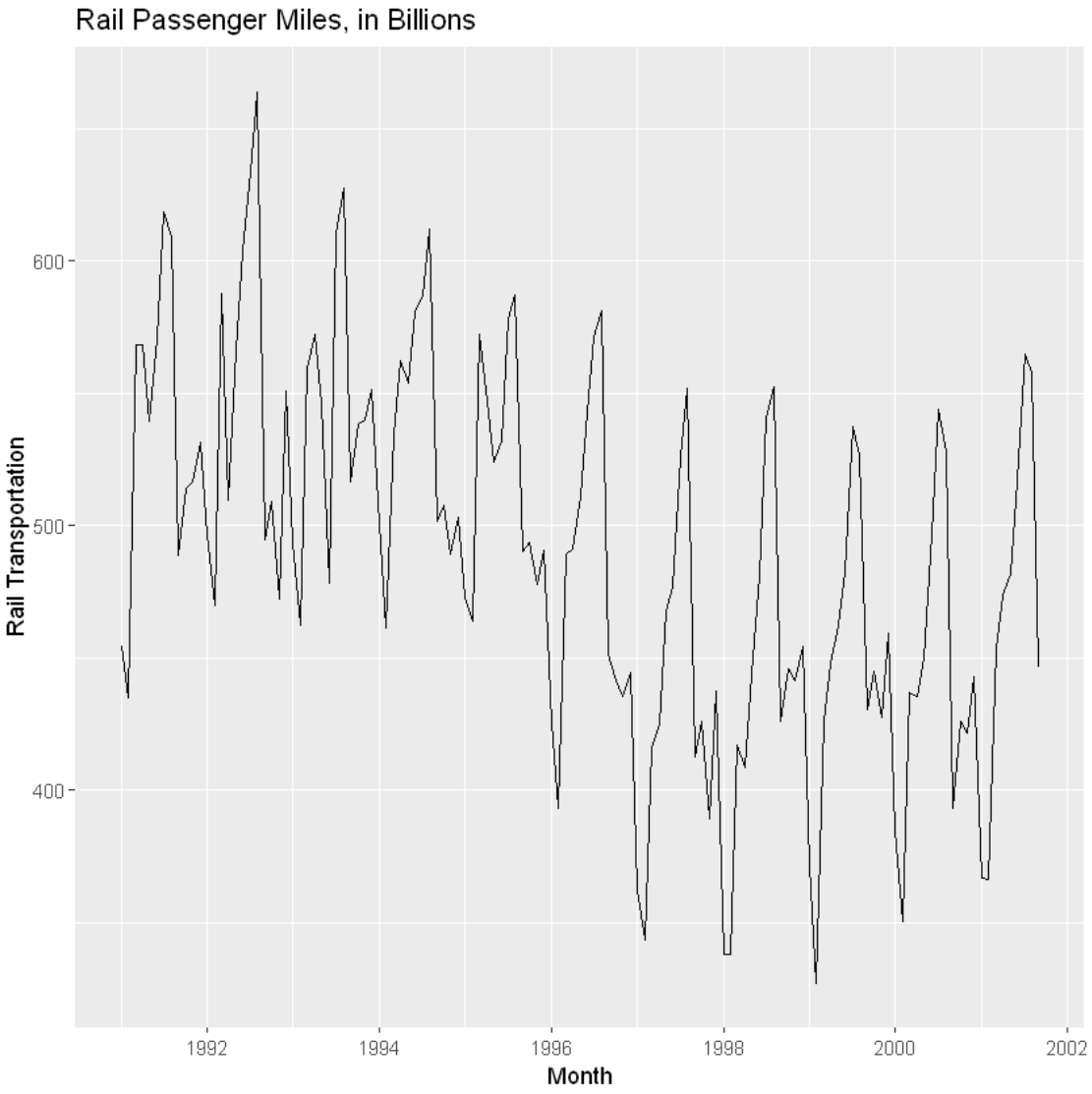
	Month <chr>	Air <int>	Rail <int>	Car <dbl>
1	01/01/1990	35153577	454115779	163.28
2	01/02/1990	32965187	435086002	153.25

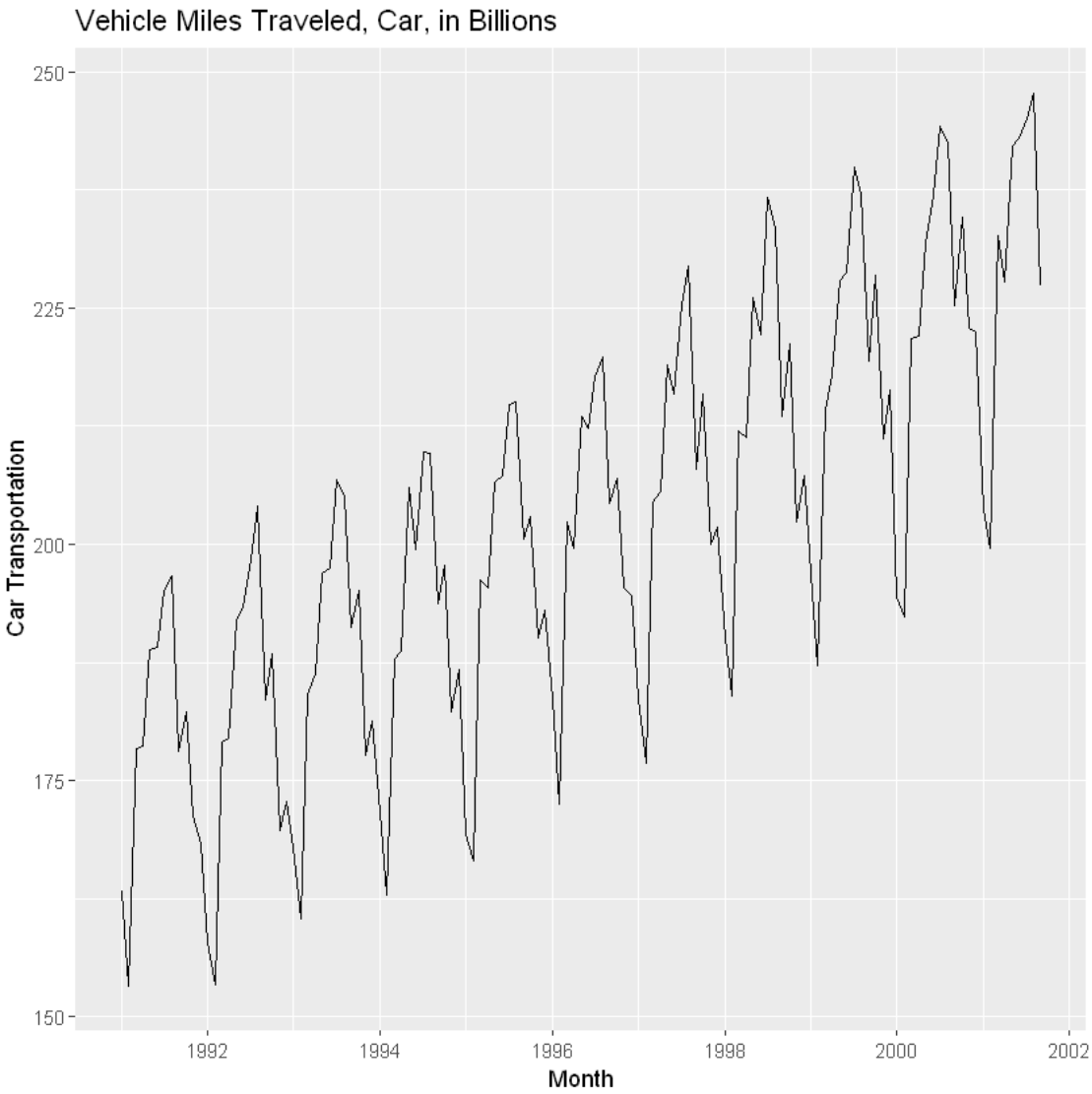
[170]:
```
n = which(df$Month == "01/09/2001") # will be used for data splitting
```

[15]:
```
# create a ts for air, rail, and car, including september because the date is␣
↪september 1st of 2001
air.ts <-ts(df$Air/1000000, start = c(1991,1), end = c(2001, 9),frequency = 12)
rail.ts <-ts(df$Rail/1000000, start = c(1991,1), end = c(2001,9), frequency =␣
↪12)
car.ts <-ts(df$Car, start = c(1991,1),end = c(2001,9), frequency = 12)
```

[17]:
```
options(scipen=999)
autoplot(air.ts, xlab = "Month", ylab = "Air Transportation",
        main = "Actual Airline Revenue Passenger Miles, in Billions")
autoplot(rail.ts, xlab = "Month", ylab = "Rail Transportation",
        main = "Rail Passenger Miles, in Billions") # is rail also in billions?
autoplot(car.ts, xlab = "Month", ylab = "Car Transportation",
        main = "Vehicle Miles Traveled, Car, in Billions")
```




Rail Passenger Miles, in Billions
Rail Transportation
600
500
400
1992
1994
1996
1998
2000
2002
Month



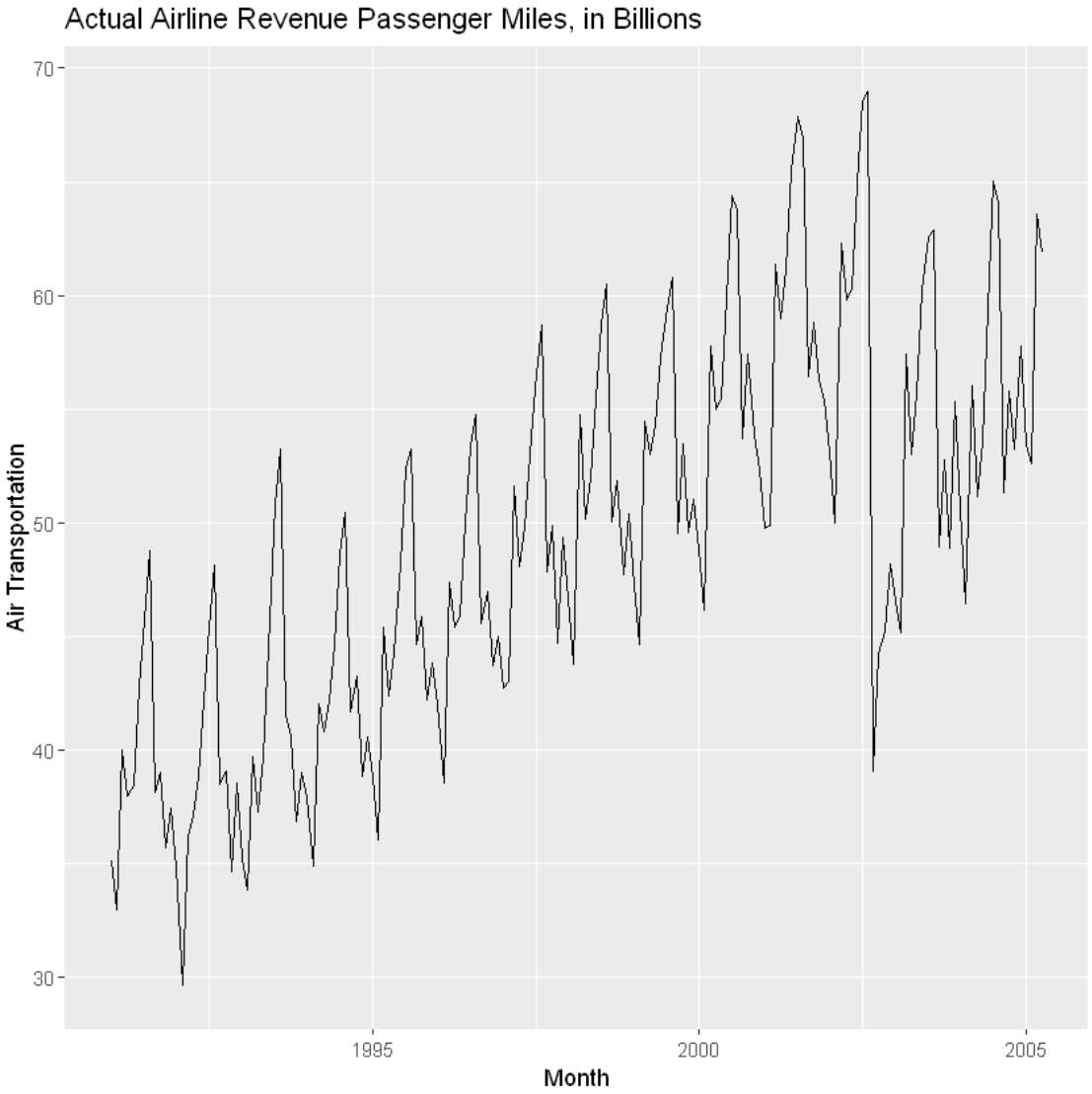
i. What time series components appear from the plot?

It seems that seasonality is present in all three of the plots due to the presence of a cyclical behavior as well as trend.

ii. What type of trend appears? Change the scale of the series, add trendlines and suppress seasonality to better visualize the trend pattern.

There is an overall upward trend in the growing number of people traveling in air and car transportation. For rail transporation, the trend seems to resemble a polynomial behavior in the rail transportation. If we further inspect the air travel, we can observe that the air travel was affected by 9/11 in approximately the year of 2002 (Figure is shown below)

[19]:
```
airfull.ts <-ts(df$Air/1000000, start = c(1991,1),frequency = 12)
autoplot(airfull.ts, xlab = "Month", ylab = "Air Transportation",
        main = "Actual Airline Revenue Passenger Miles, in Billions")
```



I changed the scale of the series in the beginning, so I will add trendlines and suppress seasonality for each type of transportation

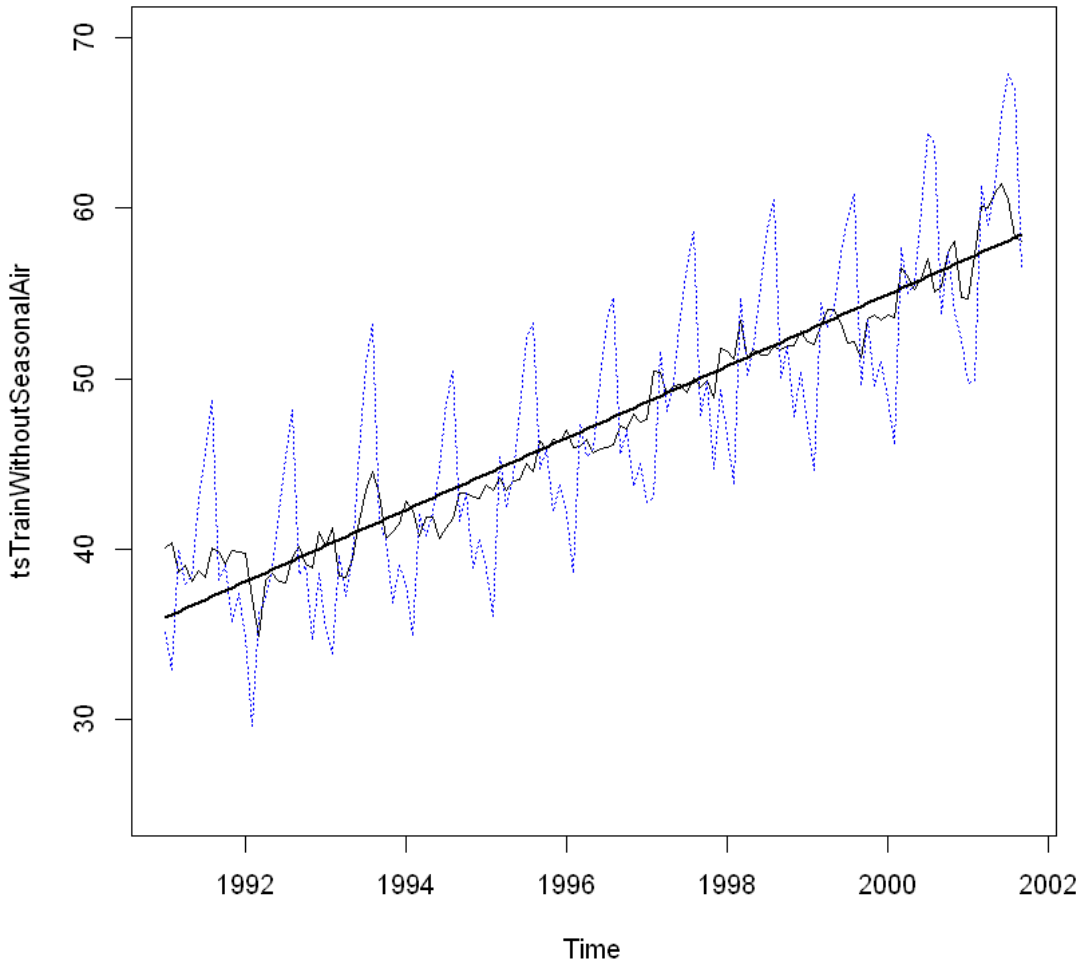
[ ]:
```
# Original Series
air.ts <-ts(df$Air/1000000, start = c(1991,1), end = c(2001, 9),frequency = 12)
rail.ts <-ts(df$Rail/1000000, start = c(1991,1), end = c(2001,9), frequency =␣
↪12)
car.ts <-ts(df$Car, start = c(1991,1),end = c(2001,9), frequency = 12)
```

[33]:
```
stlTrainAir = stl(air.ts,s.window="periodic")
stlTrainRail = stl(rail.ts,s.window="periodic")
stlTrainCar = stl(car.ts,s.window="periodic")
```

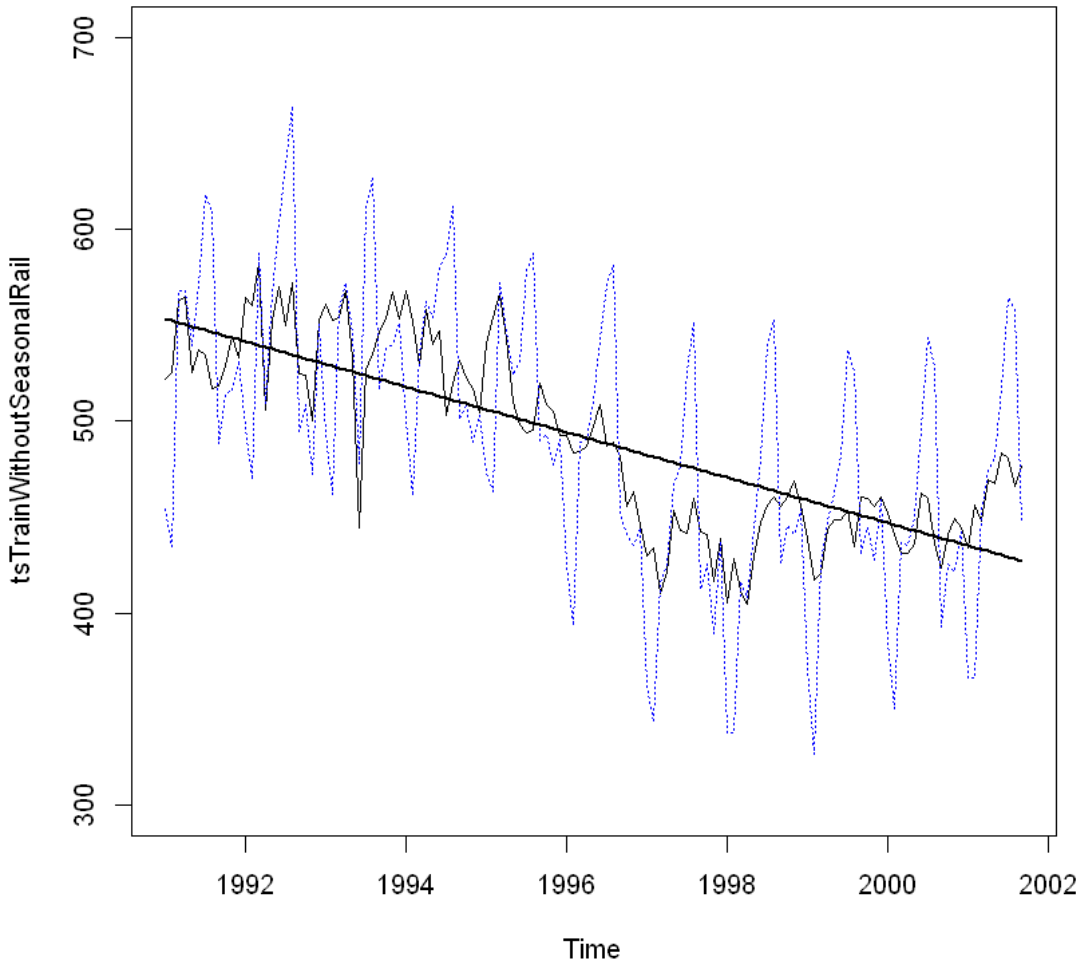
[36]:
```
#stlTrainAir - includes season, trend, and remainder
```

[38]:
```
tsTrainWithoutSeasonalAir = stlTrainAir$time.series[,2] + stlTrainAir$time.
 ↪series[,3]
tsTrainWithoutSeasonalRail = stlTrainRail$time.series[,2] + stlTrainRail$time.
 ↪series[,3]
tsTrainWithoutSeasonalCar = stlTrainCar$time.series[,2] + stlTrainCar$time.
 ↪series[,3]
```

[40]:
```
# air plot
plot(tsTrainWithoutSeasonalAir, ylim = c(25, 70)) # seasonality supressed
lines(air.ts, col = "blue", lty = 3) # actual
air.tsml <- tslm(air.ts ~ trend)
lines(air.tsml$fitted, lwd = 2) # trendline
```

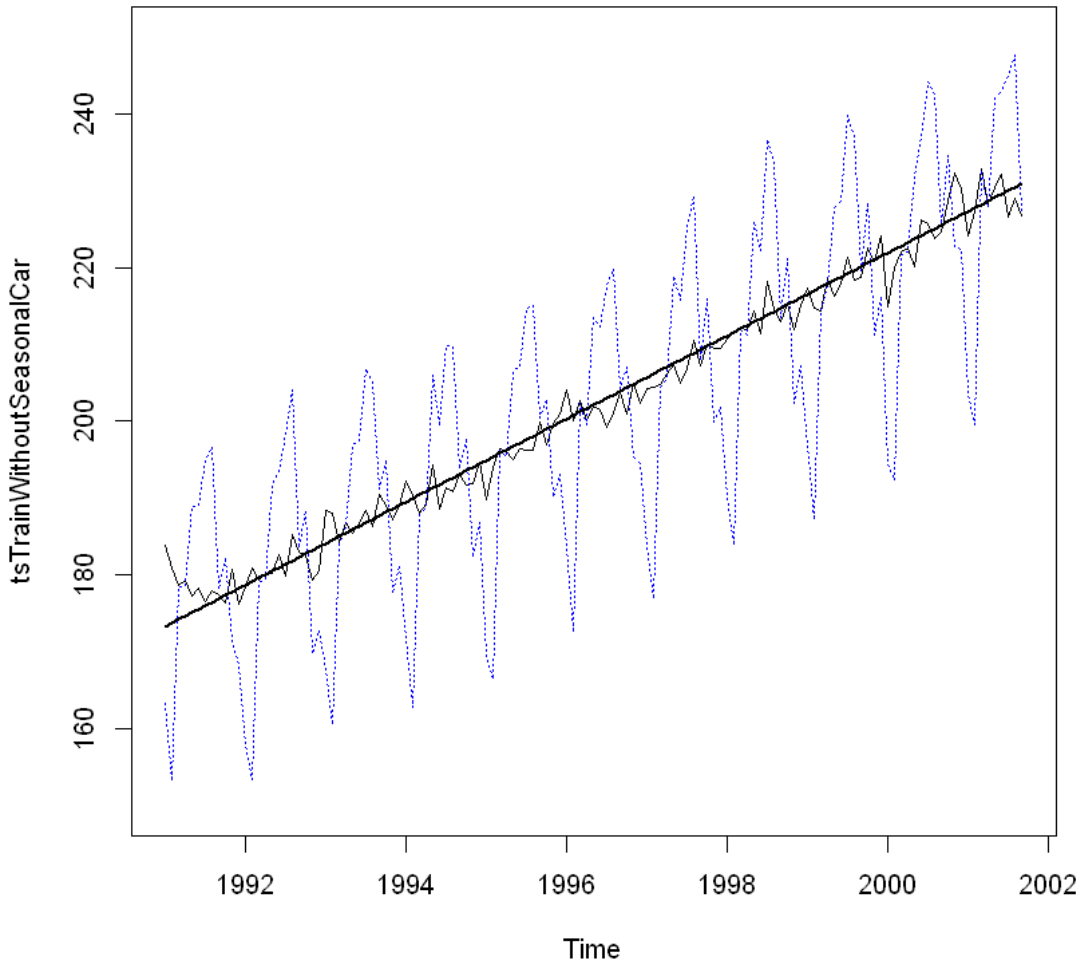


[41]:
```
# rail plot
plot(tsTrainWithoutSeasonalRail, ylim = c(300, 700)) # seasonality supressed
lines(rail.ts, col = "blue", lty = 3) # actual
rail.tsml <- tslm(rail.ts ~ trend)
lines(rail.tsml$fitted, lwd = 2) # trendline
```



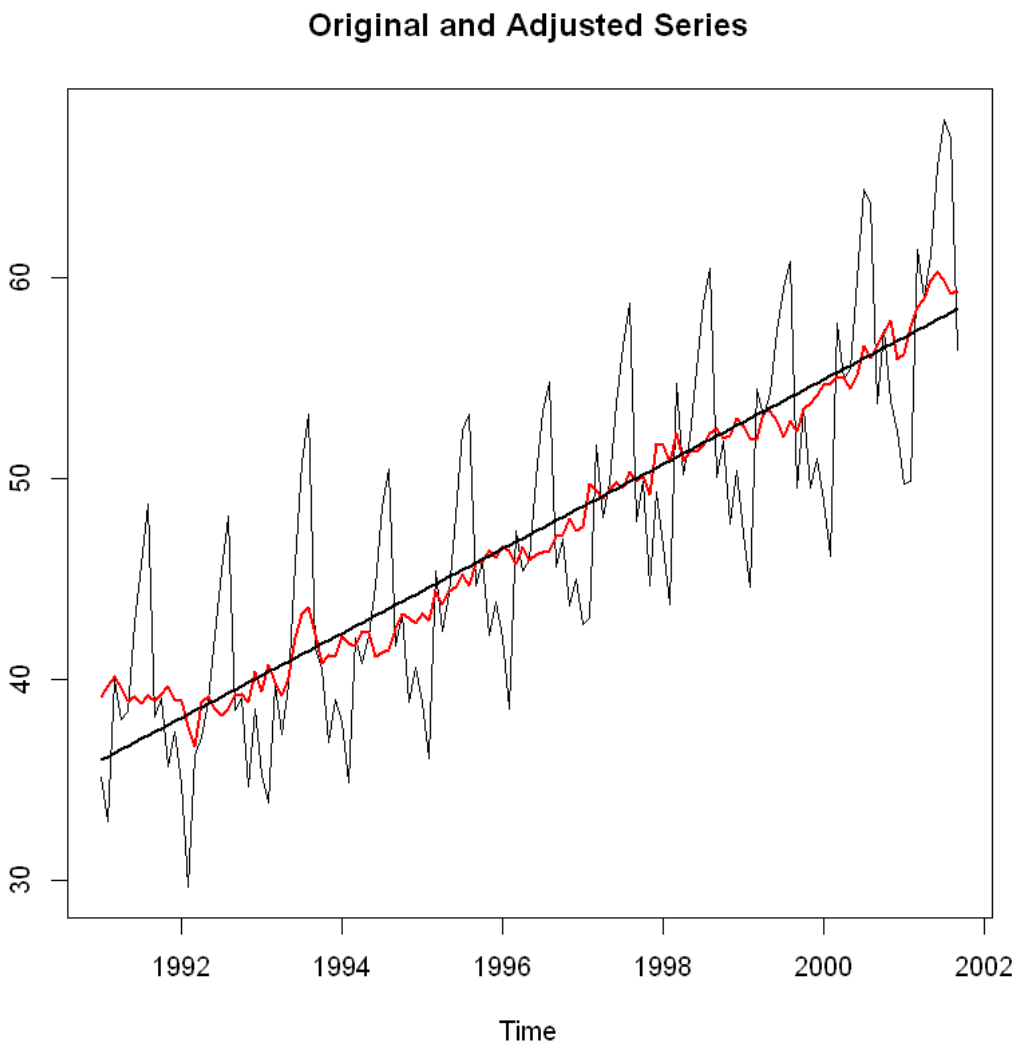
After removing seasonality, it seems that the trend is actually declining

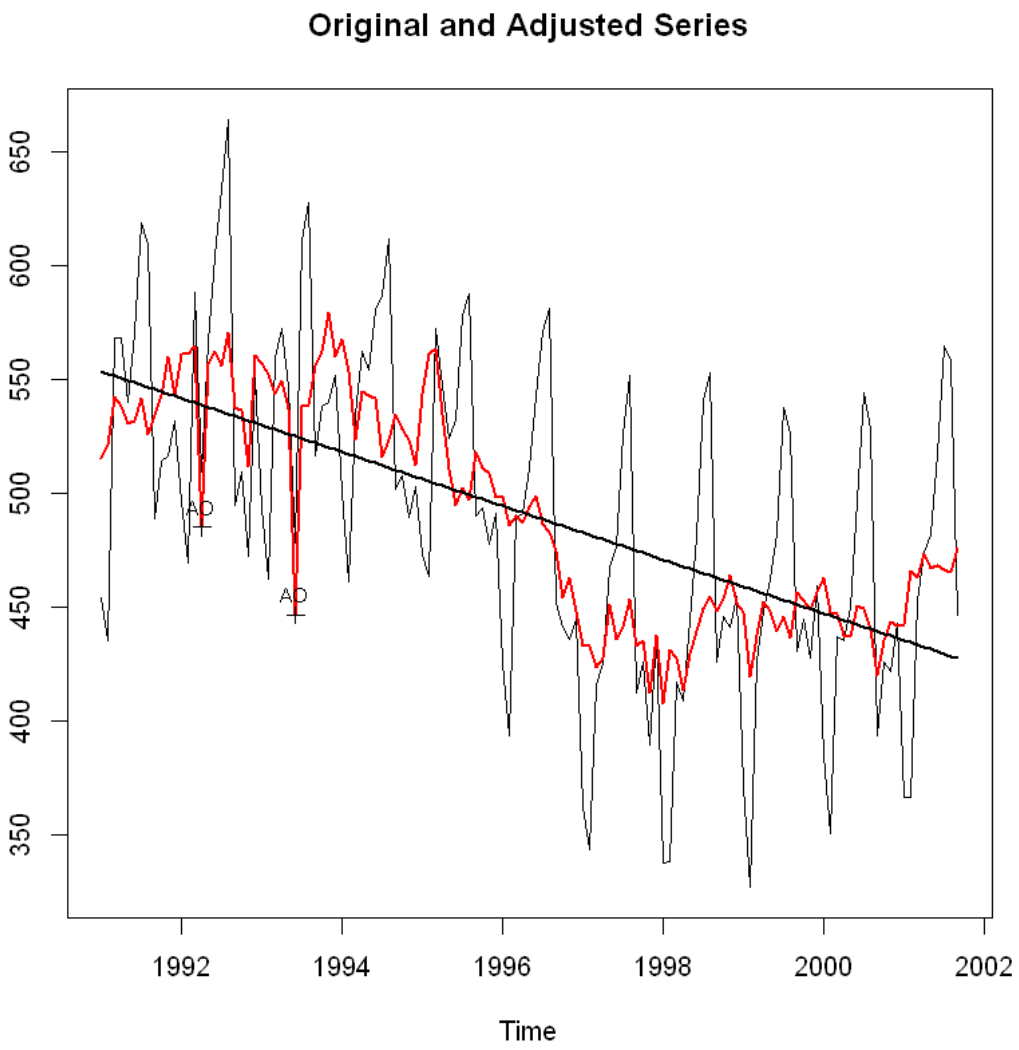
[42]:
```
# car plot
plot(tsTrainWithoutSeasonalCar, ylim = c(150, 250)) # seasonality supressed
lines(car.ts, col = "blue", lty = 3) # actual
car.tsml <- tslm(car.ts ~ trend)
lines(car.tsml$fitted, lwd = 2) # trendline
```

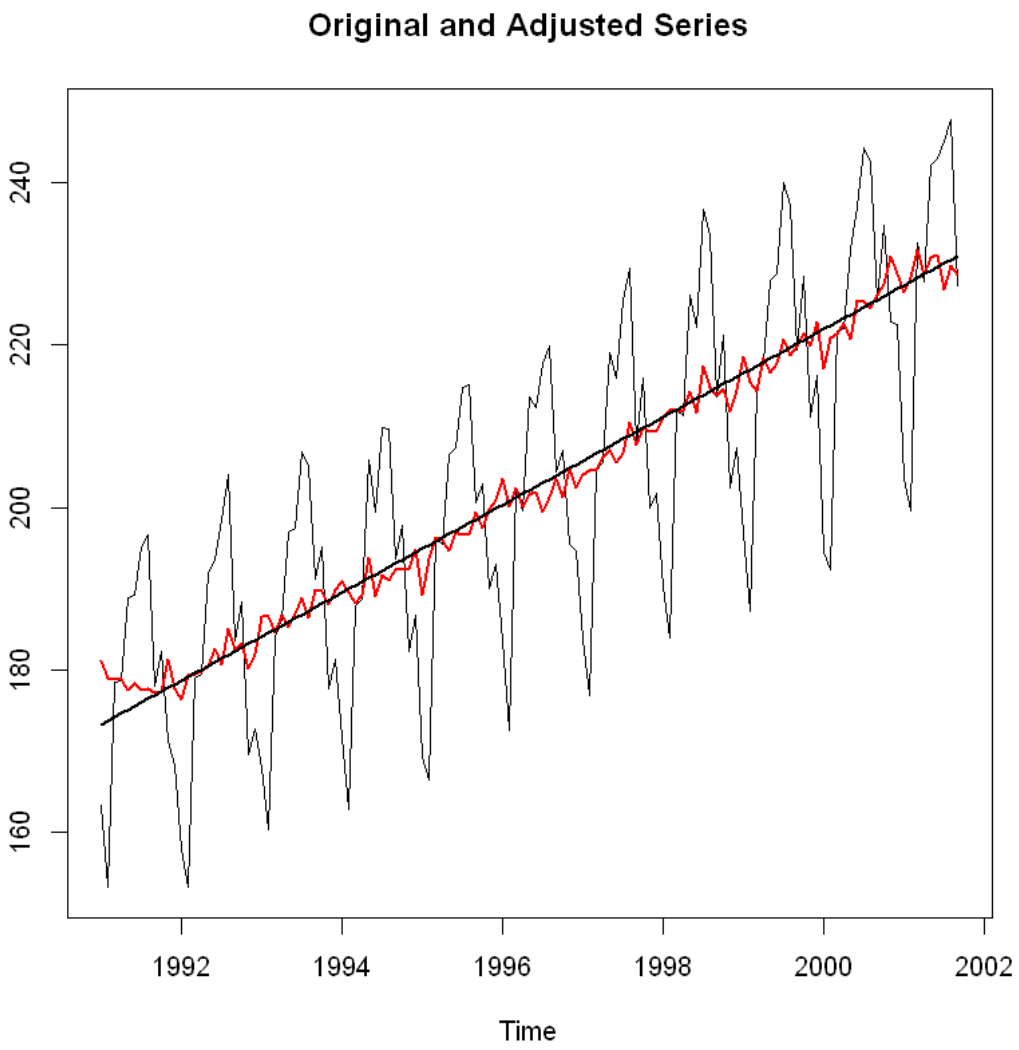


[45]:
```
#install.packages("seasonal")
```

[52]:
```
# Alternative way of drawing graphs with removing seasonality
library(seasonal)
seasonaladjustedAir = seas(air.ts)
plot(seasonaladjustedAir)
lines(air.tsml$fitted, lwd = 2) # trendline
seasonaladjustedRail = seas(rail.ts)
plot(seasonaladjustedRail)
lines(rail.tsml$fitted, lwd = 2) # trendline
seasonaladjustedCar = seas(car.ts)
plot(seasonaladjustedCar)
lines(car.tsml$fitted, lwd = 2) # trendline
```


Original and Adjusted Series
60
50
40
30
1992
1994
1996
1998
2000
2002
Time


Original and Adjusted Series
650
600
550
500
450
400
350
AO
AO
1992
1994
1996
1998
2000
2002
Time



Question 16.5

Canadian Manufacturing Workers Workhours. The time plot in Figure 16.5 describes the average annual number of weekly hours spent by Canadian manufacturing workers (data are available in CanadianWorkHours.csv—thanks to Ken Black for the data).

a. Reproduce the time plot.

[54]:
```
library(forecast)
library(ggplot2)
```

[55]:
```
df <-read.csv("CanadianWorkHours.csv", stringsAsFactors = FALSE)
```

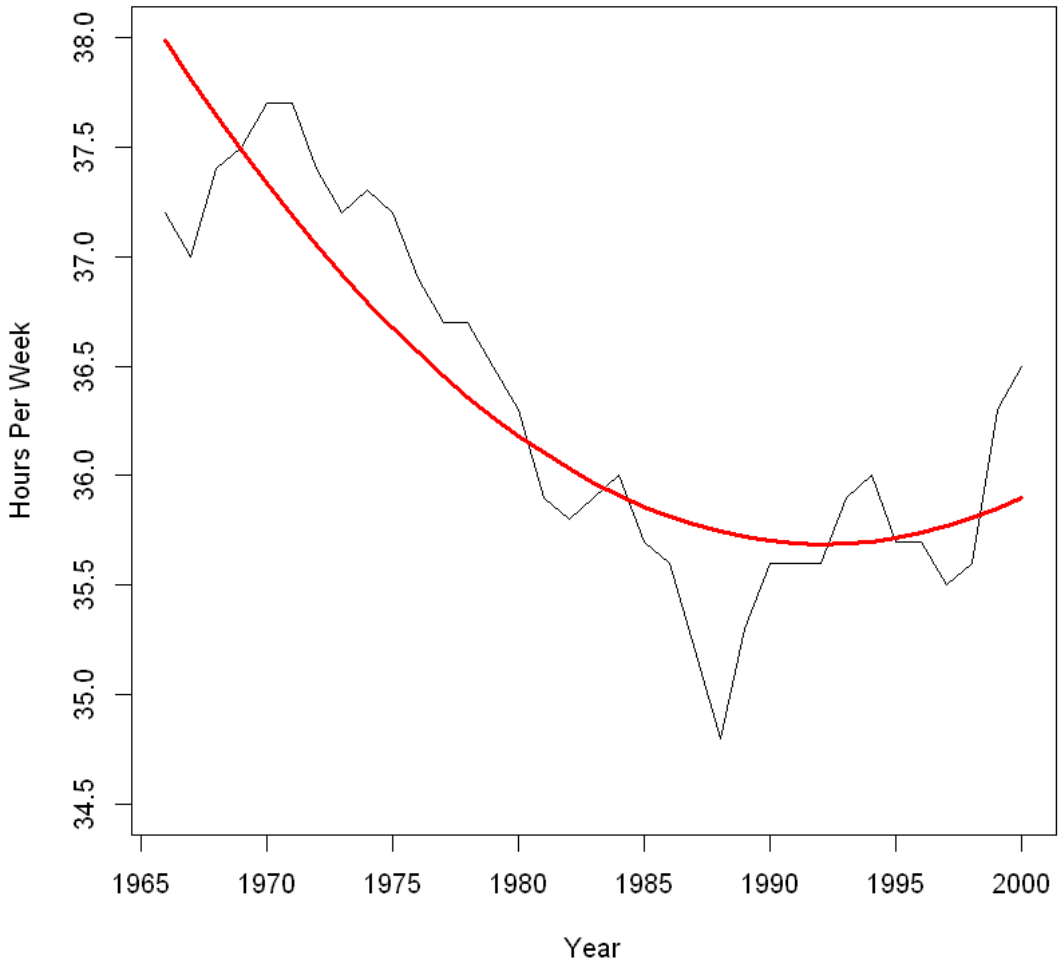
[56]:
```
tail(df,2)
```

A data.frame: 2 × 2

	Year <int>	Hours <dbl>
34	1999	36.3
35	2000	36.5

[57]:
```
hours.ts <- ts(df$Hours, start = c(1966),
               end = c(2000), freq = 1)
```

[64]:
```
plot(hours.ts, xlab = "Year", ylab = "Hours Per Week", ylim = c(34.5,38.0))
quadlm <- tslm(hours.ts ~ poly(trend, 2, raw=TRUE)) # show the trend
lines(quadlm$fitted, col="red", lwd = 3)
```



b. Which of the four components (level, trend, seasonality, noise) seem to be present in this series?

Seasonality does not appear to be present in this series because there is no cyclical pattern observed. It may be possible because data is not granular enough. Zooming in on the data will be able to show seasonality in the data. The trend seems to be a polynomial (i.e. quadratic) There is a possibility for noise existence in the year of 1988 Level - ???

Question 16.7

The file ShampooSales.csv contains data on the monthly sales of a certain shampoo over a 3-year period. Source: Hyndman, R.J., Time Series Data Library, http://data.is/TSDLdemo. Accessed on 07/25/15).

[112]:
```
df <-read.csv("ShampooSalesup.csv",  stringsAsFactors = FALSE)
```

[113]:
```
tail(df,2)
```

A data.frame: 2 × 2

	Month <chr>	Shampoo.Sales <dbl>
35	01/11/1997	581.3
36	01/12/1997	646.9

[114]:
```
colnames(df)
```

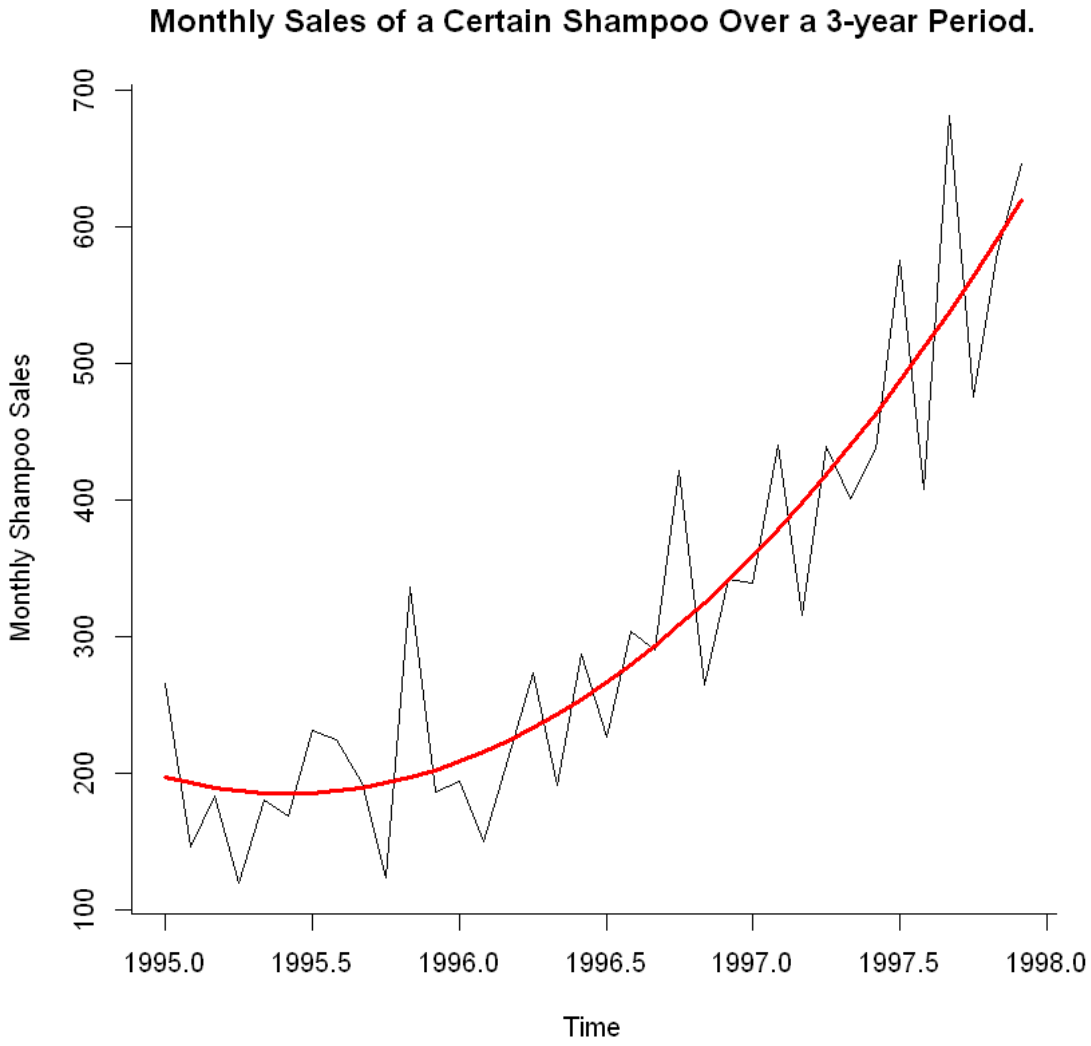
1. 'Month' 2. 'Shampoo.Sales'

[115]:
```
names(df)[2] <- "Sales"
```

a. Create a well-formatted time plot of the data.

[133]:
```
sales.ts <- ts(df$Sales, start = c(1995,1),end = c(1997, 12), frequency = 12)
lm <- tslm(sales.ts ~ trend + I(trend^2))
```

[134]:
```
plot(sales.ts, ylab = "Monthly Shampoo Sales", bty = "l", main = "Monthly Sales␣
↪of a Certain Shampoo Over a 3-year Period.")
lines(lm$fitted, col="red", lwd = 3)
```



b. Which of the four components (level, trend, seasonality, noise) seem to be present in this series?

[135]:
```
decompose(sales.ts)
```

```
$x
       Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep   Oct   Nov   Dec
1995 266.0 145.9 183.1 119.3 180.3 168.5 231.8 224.5 192.8 122.9 336.5 185.9
1996 194.3 149.5 210.1 273.3 191.4 287.0 226.0 303.6 289.9 421.6 264.5 342.3
1997 339.7 440.4 315.9 439.3 401.3 437.4 575.5 407.6 682.0 475.3 581.3 646.9

$seasonal
            Jan        Feb        Mar        Apr        May        Jun
1995 -19.193924  -2.218924 -48.175174  27.591493 -44.800174   6.345660
```

```
1996 -19.193924  -2.218924 -48.175174  27.591493 -44.800174   6.345660
1997 -19.193924  -2.218924 -48.175174  27.591493 -44.800174   6.345660
           Jul        Aug        Sep        Oct        Nov        Dec
1995   2.951910  30.431076  -1.171007  20.295660  37.274826  -9.331424
1996   2.951910  30.431076  -1.171007  20.295660  37.274826  -9.331424
1997   2.951910  30.431076  -1.171007  20.295660  37.274826  -9.331424

$trend
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
1995       NA       NA       NA       NA       NA       NA 193.4708 190.6333
1996 216.4250 219.4792 226.8208 243.3125 252.7583 256.2750 268.8500 287.0292
1997 366.3875 385.2833 405.9542 424.5292 439.9667 465.8583       NA       NA
          Sep      Oct      Nov      Dec
1995 191.9083 199.4500 206.3292 211.7292
1996 303.5583 314.8833 330.5458 345.5583
1997       NA       NA       NA       NA

$random
             Jan         Feb         Mar         Apr         May         Jun
1995          NA          NA          NA          NA          NA          NA
1996   -2.931076  -67.760243   31.454340    2.396007  -16.558160   24.379340
1997   -7.493576   57.335590  -41.878993  -12.820660    6.133507  -34.803993
             Jul         Aug         Sep         Oct         Nov         Dec
1995   35.377257    3.435590    2.062674  -96.845660   92.896007  -16.497743
1996  -45.801910  -13.860243  -12.487326   86.421007 -103.320660    6.073090
1997          NA          NA          NA          NA          NA          NA

$figure
 [1] -19.193924  -2.218924 -48.175174  27.591493 -44.800174   6.345660
 [7]   2.951910  30.431076  -1.171007  20.295660  37.274826  -9.331424

$type
[1] "additive"

attr(,"class")
[1] "decomposed.ts"
```

From the decompose function, it seems that all 4 components are present (where random = noise and x = level) From the plot, it seems that there is an overall upward quadratic trend in the Shampoo sales. The overall level is about 400 units of shampoo.

c. Do you expect to see seasonality in sales of shampoo?

Even though shampoo is an item that should be purchased every month because of its frequent use, there are noticable dips and spikes in the Shampoo sales. One possible explanation for such changes is due to the fact that, on average, a shampoo bottle lasts 2-3 months.

d. If the goal is forecasting sales in future months, which of the following steps should be taken?

• Partition the data into training and validation sets: Yes. Partitioning the data will enable development of an accurate forecasting model that is nor overfitted

• Tweak the model parameters to obtain good fit to the validation data: ???

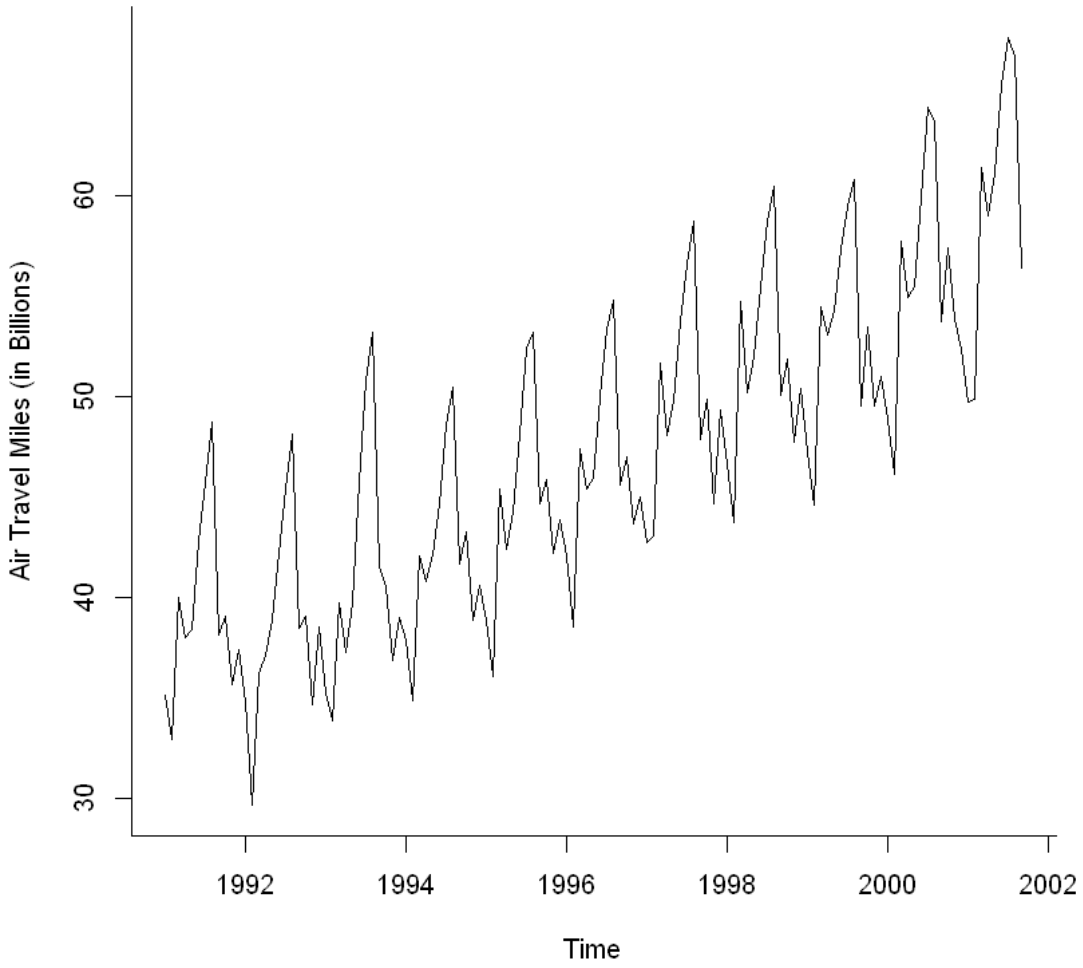
• Look at MAPE and RMSE values for the training set: No. The MAPE and RSME values we are more interested is in the validation set as opposed to training.

• Look at MAPE and RMSE values for the validation set: Yes. MAPE gives a percentage score of how predictions, on average, deviate from the actual values. RMSE, which is an equivalent to a standard error of estimate in a linear regression and is computed on the validation data set. Those measures are normally used to compare models and assess their accuracy.

Chapter 17

a. Plot the pre-event AIR time series. What time series components appear?

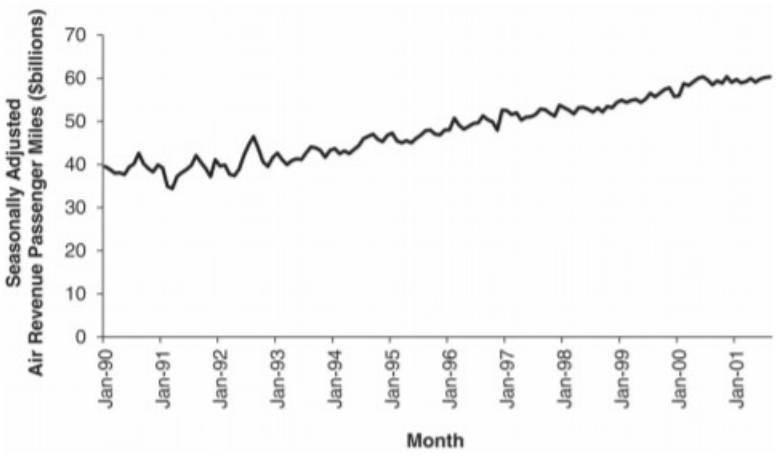
[136]:
```
air.ts.pre <- air.ts
plot(air.ts.pre, ylab = "Air Travel Miles (in Billions)",
    ylim = c(min(air.ts), max(air.ts)), bty = "l")
```



Cyclical seasonality occurs along with an overall upward trend

b. Figure 17.11 shows a time plot of the seasonally adjusted pre-September-11 AIR series. Which of the following methods would be adequate for forecasting the series shown in the figure?

[244]:
```
library("IRdisplay")
display_png(file="171.jpg")
```



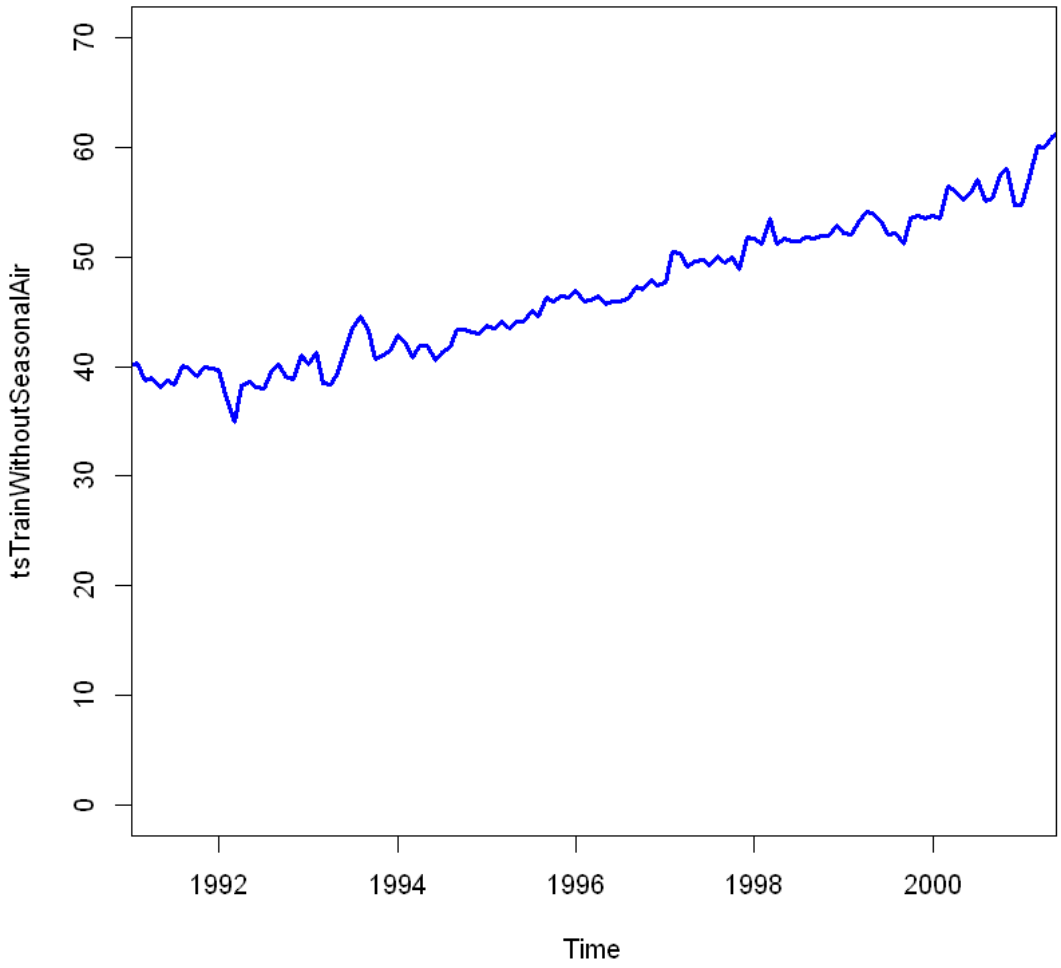
Linear regression model with trend would be adequate to use in this case since seasonally adjusted means removal of a seasonal component of a time series

[ ]:
```
c. Specify a linear regression model for the AIR series that would produce a␣
↪seasonally adjusted series similar to the one shown in Figure 17.11, with␣
↪multiplicative
seasonality. What is the outcome variable? What are the predictors?
```

I will be using tsTrainWithoutSeasonalAir model, which is previously defined. It will just include the trend, noise, and level as predictors and the outcome variable still will be the time

d. Run the regression model from (c). Remember to use only pre-event data.

[138]:
```
plot(tsTrainWithoutSeasonalAir, ylim = c(0, 70), xlim = c(1991.4, 2001), lwd =␣
↪3, col = "blue") # seasonality supressed
```



i. What can we learn from the statistical insignificance of the coefficients for October and September?

Statistical insignificance means that the results produced in September and October are approximately the same

ii. The actual value of AIR (air revenue passenger miles) in January 1990 was 35.153577 billion. What is the residual for this month, using the regression model? Report the residual in terms of air revenue passenger miles.

[145]:
```
##############################################################################################
```

[144]:
```
# did not finish######
```

Question 17.2

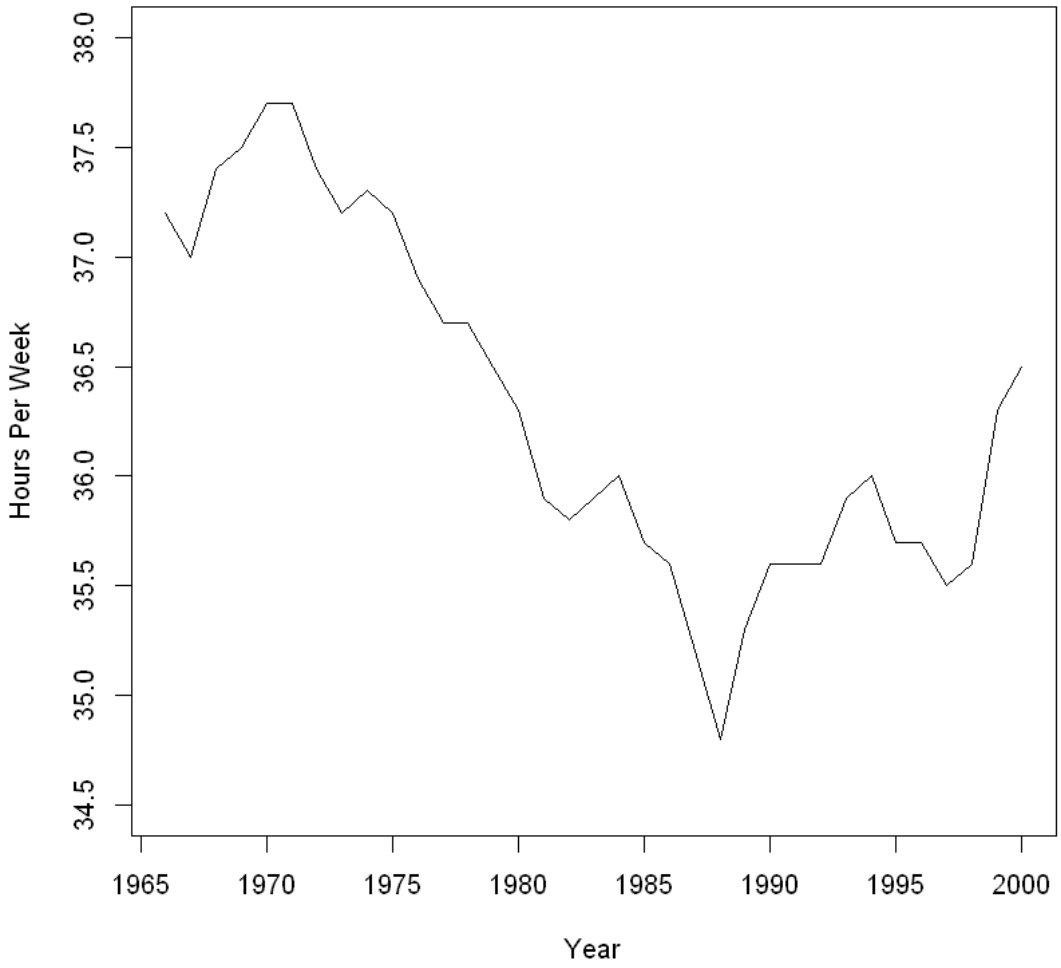
Analysis of Canadian Manufacturing Workers Workhours. The time plot in Figure 17.12 describes the average annual number of weekly hours spent by Canadian manufacturing workers (data are available in CanadianWorkHours.csv, data courtesy of Ken Black).

a. Which of the following regression models would fit the series best? (Choose one.)

[260]:
```
display_png(file="172.jpg")
```

Quadratic trand model would fit the best. There is no seasonality in the plot. As a proof, decompose function will return an error

[268]:
```
df <-read.csv("CanadianWorkHours.csv", stringsAsFactors = FALSE)
hours.ts <- ts(df$Hours, start = c(1966),
               end = c(2000), freq = 1)
plot(hours.ts, xlab = "Year", ylab = "Hours Per Week", ylim = c(34.5,38.0))
```



[267]:
```
decompose(hours.ts)
```

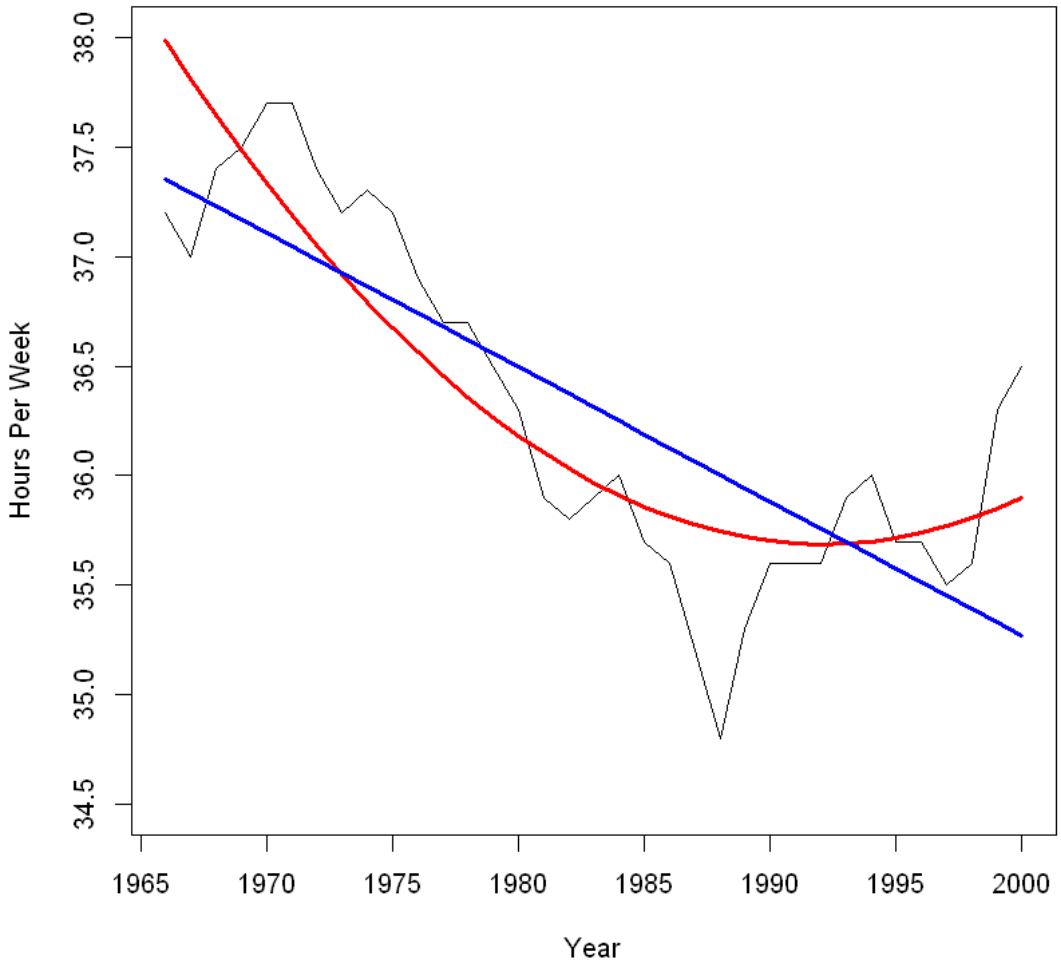
```
    Error in decompose(hours.ts): time series has no or less than 2 periods
Traceback:


    1. decompose(hours.ts)

    2. stop("time series has no or less than 2 periods")
```

Any option with seasonality disappears. Now, the trend is not linear because hours per week descrease and then start increasing from 1987. Proof is in the code.

[277]:
```
quadlm <- tslm(hours.ts ~ poly(trend, 2, raw=TRUE))
linear <- tslm(hours.ts ~ poly(trend, 1, raw=TRUE))
plot(hours.ts, xlab = "Year", ylab = "Hours Per Week", ylim = c(34.5,38.0))
lines(quadlm$fitted, col="red", lwd = 3)
lines(linear$fitted, col="blue", lwd = 3)
```



b. If we computed the autocorrelation of this series, would the lag-1 autocorrelation exhibit negative, positive, or no autocorrelation? How can you see this from the plot?

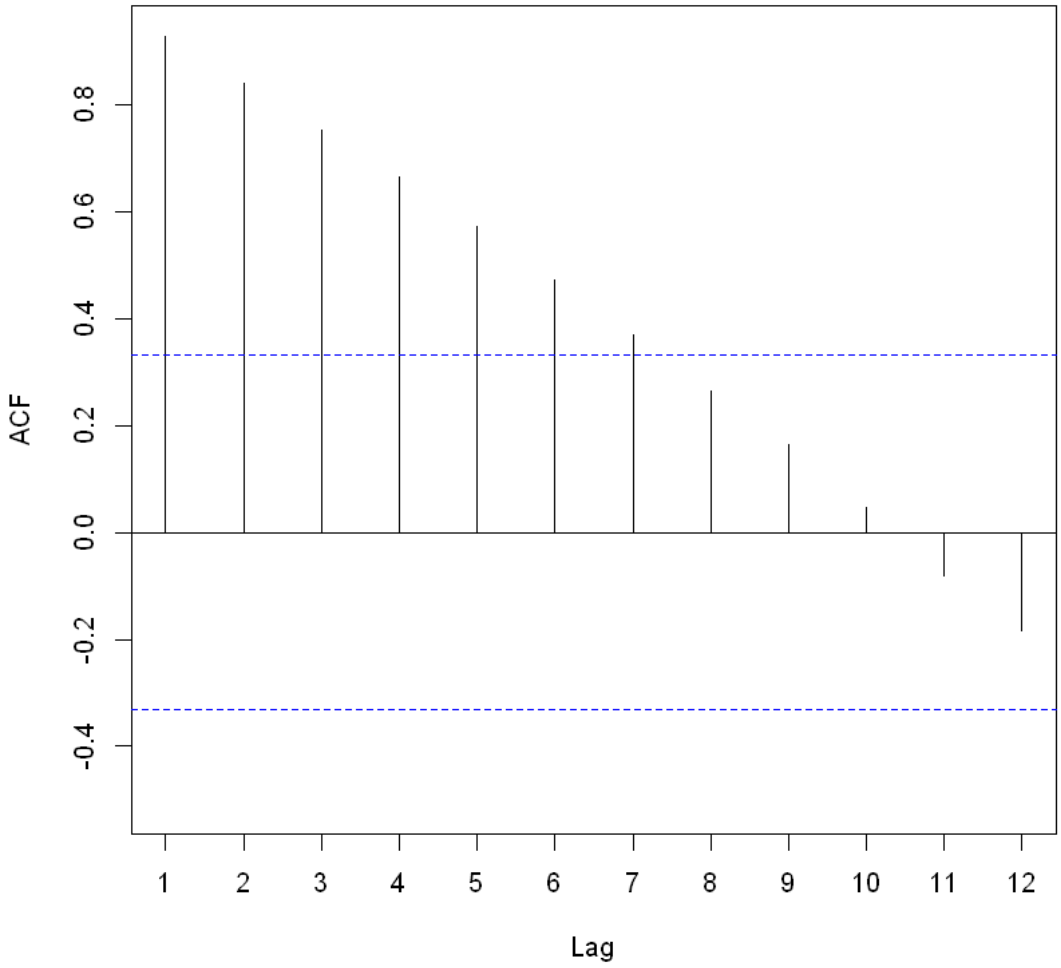
to compute the lag-1 autocorrelation, which measures the linear relationship between values in consecutive time periods

Taking the time series graph into consideration, it is possible to see that the ?

c. Compute the autocorrelation of the series and produce an ACF plot. Verify your answer to

the previous question.

[269]:
```
Acf(hours.ts, lag.max = 12, main = "")
```



lag 1 executes a positive autocorrelation, which can be seen from the autocorrelation plot above

Question 17.5

The time plot in Figure 17.15 describes actual quarterly sales for a department store over a 6-year period (data are available in DepartmentStoreSales.csv, data courtesy of Chris Albright).

a. The forecaster decided that there is an exponential trend in the series. In order to fit a regression-based model that accounts for this trend, which of the following operations must be performed?

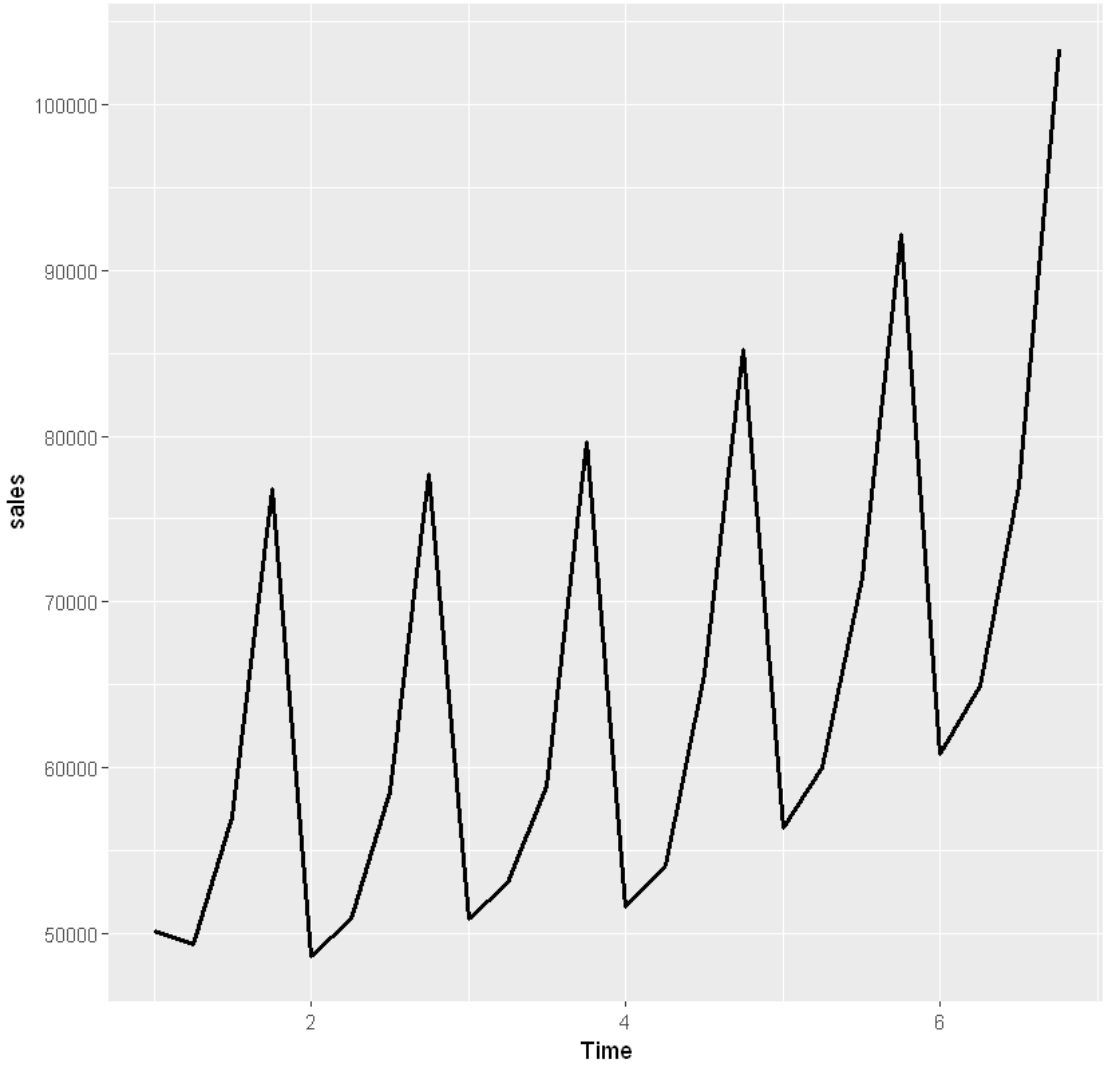
Take the log of sales because sales is the dependent variable?

b. Fit a regression model with an exponential trend and seasonality, using the first 20 quarters as the training data (remember to first partition the series into training and validation series).

[279]:
```
df <- read.csv("DepartmentStoreSales.csv")
```

[300]:
```
sales <- ts(df$Sales, frequency=4)
```

[302]:
```
autoplot(sales, lwd = 1)
```



[286]:
```
nTrain <- 20
nValid <- length(df$Quarter) - nTrain
```

[306]:
```
salesTrain <- window(sales, end=c(1, nTrain))
salesValid <- window(sales, start=c(1,nTrain+nValid))
```

[307]:
```
salesExpo <- tslm(salesTrain ~ trend + season, lambda=0)
```

[309]:
```
summary(salesExpo)
```

```
Call:
tslm(formula = salesTrain ~ trend + season, lambda = 0)

Residuals:
      Min        1Q    Median        3Q       Max
-0.053524 -0.013199 -0.004527  0.014387  0.062681

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 10.748945   0.018725 574.057 < 0.0000000000000002 ***
trend        0.011088   0.001295   8.561      0.0000003702488 ***
season2      0.024956   0.020764   1.202                0.248
season3      0.165343   0.020884   7.917      0.0000009788403 ***
season4      0.433746   0.021084  20.572      0.0000000000021 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03277 on 15 degrees of freedom
Multiple R-squared:      1,Adjusted R-squared:      1
F-statistic: 7.63e+11 on 4 and 15 DF,  p-value: < 0.00000000000000022
```

c. A partial output is shown in Table 17.10. From the output, after adjusting for trend, are Q2 average sales higher, lower, or approximately equal to the average Q1 sales?

The high value of p-value of .248 for the 2nd quarter means that the values for the first and second quarters are approximately the same

d. Use this model to forecast sales in quarters 21 and 22.

[310]:
```
salesForecast <- forecast(salesExpo, h=2)
```

[311]:
```
salesForecast
```

```
     Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
6 Q1       58793.71 55790.19 61958.92 54090.84 63905.46
6 Q2       60951.51 57837.76 64232.89 56076.04 66250.87
```

e. The plots in Figure 17.16 describe the fit (top) and forecast errors (bottom) from this regression model.

f. Recreate these plots.

[ ]: